TAMADDUN JURNAL BAHASA, SASTRA DAN BUDAYA LIFE
ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

What Gets You a High Band Score? A Conversation Analysis of the IELTS Speaking Test

Adamry Muis[1]*, Ashabul Kahfi Susanto[2]
Email: 044233855@ecampus.ut.ac.id, *ashabul_susanto@unm.ac.id*
[1]Universitas Terbuka, Indonesia
[2]Universitas Negeri Makassar, Indonesia

Received: 14 July 2026 **Accepted**: 30 July 2026 **Published**: 21 August 2026

Abstract

The relationship between the linguistic features that candidates produce during oral proficiency interviews and the band scores they are subsequently awarded remains a central, yet still under-specified, concern in language testing research. Drawing on Conversation Analysis (CA), this article examines the interactional organisation of the IELTS Speaking Test (IST) and re-examines how candidate talk at different score levels differs in its topical, sequential, and lexico-grammatical properties. Sixteen extracts drawn from publicly available IELTS test recordings were transcribed using standard CA conventions and analysed turn-by-turn for topical coherence, engagement with the topic, repair practice, identity construction, colloquial delivery, lexical choice, hesitation phenomena, and syntactic complexity. The analysis shows that higher-scoring candidates display topic elaboration beyond minimal response, deploy repair and clarification requests as resources for maintaining fluency, construct favourable interactional identities, and produce syntactically complex, low-dysfluency talk, whereas lower-scoring candidates orient minimally to topic development, exhibit extended hesitation and self-repair trouble, and produce comparatively simple morphosyntax. These patterns are discussed in relation to the official IELTS band descriptors and to the wider construct of interactional competence, and are triangulated against recent quantitative and discourse-analytic studies of the Speaking Test. The article argues that candidate talk is not merely a symptom of an underlying score but an interactionally accomplished performance that examiners and raters orient to, with implications for rater training, task design, and the validity of oral assessment more broadly. Methodological limitations associated with a small, opportunistically sampled corpus of publicly available recordings are acknowledged, and directions for larger-scale, ethically sourced CA-based validation research are proposed.

Keywords: *Conversation analysis; IELTS Speaking Test; candidate talk; band score; interactional competence*

INTRODUCTION

High-stakes tests of second-language (L2) speaking ability now function as gatekeeping instruments for university admission, professional migration, and civic participation across the English-speaking world, and the IELTS Speaking Test (hereafter IST) is among the most widely used of these instruments (Ducasse & Brown, 2011; Dashti & Razmjoo, 2020). Millions of candidates a year encounter the IST as the final, and often most anxiety-inducing, hurdle in an otherwise paper-based or computer-based test battery, precisely because it is assessed live, individually, and in a compressed span of eleven to fourteen minutes (Issitt, 2008). The stakes

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

attached to the test - visa applications, university offers, professional accreditation - make the question of how a band score is arrived at, and what in the candidate's talk actually warrants that score, a matter of both theoretical and practical urgency (Taylor, 2000; O'Loughlin, 2002; Seedhouse, 2012).

Speaking assessment researchers have approached this question from two broad, and largely complementary, directions. The first is psychometric and discourse-analytic: researchers quantify features such as speech rate, pausing, lexical diversity, grammatical accuracy and complexity, and cohesive-device use, and correlate these with the score awarded (Brown, 2006; Iwashita & Vasquez, 2015; Dashti & Razmjoo, 2020). Brown (2006), working within the IELTS validation research programme, identified linguistic competence, fluency, pronunciation, and content as the four broad dimensions along which raters appear to differentiate candidate performance. Dashti and Razmjoo's (2020) mixed-methods analysis of 59 mock IELTS performances found that fluency and coherence were the strongest predictors of total speaking score, ahead of grammatical range and lexical resource, while pronunciation contributed comparatively little unique variance - a finding that nuances, without overturning, Brown's earlier four-way model. Iwashita and Vasquez (2015) similarly showed that discourse-competence features, particularly cohesive and coherence devices, were more consistently associated with proficiency level than any single lexico-grammatical measure taken alone.

The second broad approach, and the one adopted in this article, is interactional: rather than treating the candidate's turn as an isolated specimen of interlanguage to be measured against a checklist, Conversation Analysis (CA) treats the IST as a jointly-produced, sequentially-organised speech event, in which the examiner's talk, the institutional script, and the candidate's talk are mutually constitutive (Seedhouse, 2004; Seedhouse & Egbert, 2006; Seedhouse & Harris, 2011; Seedhouse, 2012; Seedhouse & Nakatsuhara, 2018). This tradition, most extensively developed in Seedhouse's programme of IELTS validation research and consolidated in the 137-test corpus reported by Seedhouse and Nakatsuhara (2018), argues that oral proficiency cannot be adequately understood by measuring candidate output alone, because a test candidate does not simply produce spontaneous discourse: they respond, moment by moment, to an interlocutor whose contributions are themselves institutionally constrained (Seedhouse & Harris, 2011). An emic, participant-oriented CA analysis is therefore held to be uniquely well placed to reveal precisely how "high-scoring" and "low-scoring" talk differs, because it examines the micro-detail of the interaction turn by turn rather than aggregating linguistic features across an entire performance (Seedhouse, 2012, p. 8). Douglas (1994) had earlier raised a related concern, observing that speakers can "produce qualitatively different performance and yet receive similar ratings" (p. 125) when a rating scale collapses genuinely distinct discourse phenomena into a small number of undifferentiated grammar-based bands; a fine-grained interactional lens recovers the qualitative distinctions that a holistic numerical score necessarily flattens.

Despite the theoretical appeal of the CA approach, Lazaraton (2002) observed that there had been "very little published work on the empirical relationship between candidate speech output and assigned ratings" (pp. 161-162), a gap subsequently only partially closed by Seedhouse and colleagues' mixed-method study of band descriptors and speaking features (Seedhouse et al., 2014), by the extended corpus analysis in Seedhouse and Nakatsuhara (2018), and by more recent interactional-competence-oriented scholarship (Roever & Ikeda, 2023; Youn, 2020; Dai, 2025). This persistent gap is consequential: because the official IELTS band descriptors do not explicitly categorise candidate talk or discourse features (Seedhouse et al., 2014), examiners' and raters' impressions of a "confident," "coherent," or "high-achieving" candidate are, in effect, being

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

formed by interactional phenomena that the descriptors do not name, even though those same phenomena appear to be systematically implicated in the score outcome. Identifying this relationship, as Seedhouse et al. (2014, p. 5) argue, "can provide valuable input for validation processes," and is squarely a matter of test fairness and construct validity rather than a purely descriptive linguistic exercise.

The present article revisits this relationship using data drawn from publicly available recordings of the IST, informed by an updated reading of the CA and language-testing literature. It is motivated by three considerations. First, more than a decade has passed since Seedhouse and Harris's (2011) and Seedhouse et al.'s (2014) foundational IELTS Research Reports, and the interim period has seen renewed scholarly interest in interactional competence (IC) as a testable construct in its own right (Roever & Dai, 2021; Roever & Ikeda, 2023; Dai, 2025), alongside quantitative replications that both confirm and complicate the earlier discourse-analytic findings (Dashti & Razmjoo, 2020; Iwashita & Vasquez, 2015). Second, the rapid emergence of automated and semi-automated spoken-language assessment technologies - from spoken dialogue systems to large language model-based scoring (Karatay & Xu, 2025; Ma et al., 2025; Goh & Aryadoust, 2025) - has sharpened rather than dissolved the need to understand precisely which interactional behaviours human examiners and raters are responding to, since any automated system is either trained to approximate, or explicitly designed to bypass, those same behaviours. Third, the pedagogical stakes of misunderstanding this relationship are high: teachers and candidates alike often prepare for the IST by rehearsing content and vocabulary while under-attending to the sequential and repair-related dimensions of live interaction that this article shows to be consequential.

Accordingly, this article addresses the following research question: what is the relationship between the interactional features of candidate talk - topical, sequential, and lexico-grammatical - and the band score a candidate receives in the IST? In pursuing this question, the article does not aim to produce a new predictive or psychometric model of scoring; rather, following the CA tradition of "description leading to informed action" (Richards & Seedhouse, 2005, cited in Seedhouse, 2011, p. 39), it aims to describe, with the analytic precision CA affords, how band-relevant distinctions are interactionally accomplished in real time, so that this description can subsequently inform rater training, task design, and validation research.

The article is organised as follows. Section 2 sets out the study's method, including the data source, transcription conventions, and the analytic procedure used to classify and compare high- and low-scoring extracts. Section 3 presents the findings, organised around eight interactionally grounded categories - topical coherence, topic engagement, trouble and repair, identity construction, colloquial delivery, lexical choice, hesitation, and syntactic complexity - each illustrated with authentic candidate-examiner extracts. Section 4 discusses these findings in relation to the IELTS band descriptors, the interactional-competence literature, and recent quantitative and AI-mediated replications, before considering implications for rater training, task design, and test validity, as well as the limitations of an opportunistically sampled, publicly available corpus. Section 5 offers a brief conclusion.

Before turning to the data, it is necessary to situate the IST itself within its institutional and interactional context, since the CA claim that candidate talk cannot be understood independently of the test's institutional organisation is foundational to everything that follows (Seedhouse & Harris, 2011; Seedhouse, 2012). The IST is delivered in three parts: an introductory question-answer phase in which the examiner elicits biographical information (Part 1); an extended, largely monologic turn produced by the candidate in response to a task card (Part 2); and a more abstract,

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

discussion-oriented question-answer phase that develops the Part 2 topic (Part 3) (IELTS, 2023). Across all three parts, the examiner is required to follow a pre-scripted set of prompts, a constraint that Seedhouse and Harris (2011) describe as rendering the test topic "pre-determined" and "written out in advance in script" (p. 14), such that the examiner's institutional role is closer to that of a script-following interlocutor than a conversational partner exercising free topical choice. This scripting is not incidental: it is precisely what allows the test to claim standardisation and comparability across candidates and examiners, and it is precisely what makes candidate talk - rather than examiner talk - the primary site in which individual variation, and therefore individual differences in score, can emerge (Seedhouse & Egbert, 2006). It is against this backdrop of institutional standardisation that the analysis of candidate talk reported below should be read: because the examiner's contribution is comparatively fixed, systematic differences observed between high- and low-scoring extracts can be attributed with some confidence to properties of candidate talk itself, rather than to variation in examiner behaviour.

Understanding this organisation also clarifies why a CA approach, rather than a purely linguistic or psychometric one, is analytically appropriate. Because turn-taking, adjacency-pair structure, and topic management in the IST are institutionally constrained in specific, describable ways, it becomes possible to isolate the contribution of candidate talk to the overall trajectory and "feel" of the interaction, and hence to the impression - and ultimately the score - that the examiner forms (Seedhouse, 2012). This is the analytic warrant for the study reported in the remainder of this article.

A final introductory point concerns terminology. Following Seedhouse (2012) and Seedhouse et al. (2014), this article uses "candidate talk" to refer to the full range of verbal and paraverbal conduct produced by the test-taker - including pausing, repair, prosody, and lexico-grammatical choice - rather than to lexico-grammatical output alone. This broader conception of talk is consistent with more recent formulations of interactional competence, which define L2 speaking proficiency partly in terms of a candidate's ability to co-construct meaning, manage topic, and repair trouble collaboratively with an interlocutor, rather than in terms of a static, individually-held grammatical repertoire (Roever & Dai, 2021; Youn, 2020; Dai, 2025). Read in this light, the analysis that follows treats each extract not simply as a sample of interlanguage to be scored against a checklist, but as a locally-managed interactional achievement whose success or failure is visible in, and partly constitutive of, the band score eventually awarded.

METHOD

1. Research design

This study adopts a qualitative, single-case-oriented Conversation Analysis (CA) design, situated within the institutional-discourse strand of CA applied to language testing generally (Seedhouse, 2004) and to the IELTS Speaking Test specifically (Seedhouse & Egbert, 2006; Seedhouse & Harris, 2010, 2011; Seedhouse, 2012; Seedhouse et al., 2014; Seedhouse & Nakatsuhara, 2018). CA is inductive and data-driven: rather than beginning from a predetermined set of linguistic categories, as a psychometric approach typically does, it builds analytic categories bottom-up from recurrent, observable phenomena in repeated listening to and reading of the recording and transcript (Seedhouse, 2012). This design suits a question about how score-relevant features are interactionally produced, moment by moment, by both candidate and examiner, rather than only what features correlate with score (Seedhouse & Harris, 2011).

The design is explicitly comparative, contrasting extracts associated with high band scores (8.0-9.0) against extracts associated with low-to-mid band scores (3.0-5.0), following the logic

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

established in Seedhouse and Harris (2010) and Seedhouse (2012), whereby the interactional "signature" of successful and unsuccessful talk is brought into relief through direct juxtaposition. The high/low contrast is treated as a heuristic for identifying recurrent patterns rather than an assumption that score is a linear function of any single feature.

2. Data source and sampling

The data comprise sixteen extracts of candidate-examiner interaction drawn from publicly and freely available online recordings - simulated and demonstration IELTS Speaking Test material, including recordings distributed via the official IELTS YouTube channel. This strategy follows precedent in language-testing CA research that draws on publicly released or simulated materials when access to a large corpus of live, anonymised examination recordings is unavailable to an individual researcher (cf. Seedhouse, 2012; Seedhouse et al., 2014; Seedhouse & Nakatsuhara, 2018, whose institutionally sanctioned corpora exceed 137 recordings). Each recording reports an accompanying band score, assigned either by a certified examiner acting as commentator or presented as an official IDP/British Council exemplar; these reported scores organise the comparative analysis descriptively rather than functioning as an independently re-verified psychometric outcome, a limitation addressed in Section 4.7.

Sixteen extracts were purposively selected to represent a spread of Part 1, Part 2, and Part 3 talk, a spread of topics (media, hobbies, leadership, community programmes, employment, travel, celebrity), and a spread of band scores from 3.0 to 9.0, where (a) recording quality permitted reliable transcription, (b) the candidate's turn(s) contained a tractable instance of one or more of the eight categories that emerged inductively (Section 2.3), and (c) a band score was explicitly reported. The sample is a purposive, illustrative CA corpus rather than a probability sample, consistent with the epistemic goals of single-case-sensitive interactional analysis (Seedhouse, 2004), and no claim of statistical representativeness is made.

3. Transcription and analytic procedure

Extracts were transcribed, or previously available transcripts checked and refined, using CA conventions developed by Atkinson and Heritage (1984) and adopted in IELTS-specific transcription protocols (Seedhouse & Harris, 2014; Seedhouse & Nakatsuhara, 2018). These conventions capture delivery features invisible in an orthographic transcript: timed and untimed pauses, in-breaths, pitch shifts, sound-stretching, overlap, latching, and cut-offs. Underlining marks emphasis; colons mark lengthening; a dash marks an abrupt cut-off; timed silences appear in parentheses to a tenth of a second and micro-pauses as a bracketed full stop; square brackets mark overlap onset/termination and an equals sign marks latching. "E" denotes the examiner and "C" the candidate. The sixteen extracts in Section 3 are reproduced exactly as transcribed and are not altered for this article.

Analysis followed the four-stage "data session" procedure standard in CA research (Seedhouse, 2004). First, each recording was heard repeatedly alongside its transcript without a predetermined coding scheme. Second, provisional categories were tested against the full extract set, retaining only those recurrent across extracts and clearly differentiating high- from low-scoring talk; this yielded the eight categories reported in Section 3 - topical coherence, topic engagement, trouble and repair, identity construction, colloquial delivery, lexical choice, hesitation, and syntactic complexity - refining categories identified in Seedhouse and Harris (2011) and Seedhouse (2012). Third, each extract underwent turn-by-turn sequential analysis: how a candidate turn responded to the prior examiner turn, how topic was initiated, maintained, or abandoned, how

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

trouble was repaired, and what lexico-grammatical resources were deployed. Fourth, extracts were paired for comparative presentation - typically one high- and one low-scoring extract addressing a structurally similar task - following Seedhouse and Harris (2010).

A central analytic distinction is Seedhouse's (2012, p. 11) "topic-as-script" versus "topic-as-action": because IST topics are institutionally pre-scripted, a candidate's task is to recognise the topic embedded in the question, answer it, and develop it through elaboration and evaluation. Seedhouse and Harris (2011, p. 12) specify four sub-tasks - understanding, answering, identifying topic, developing topic - which informed Sections 3.1-3.2, while the remaining six categories capture sequential, identity, and lexico-grammatical phenomena this framework does not by itself capture.

4. Ethics and trustworthiness

Because the extracts are drawn from recordings placed in the public domain for demonstration, revision, or commentary purposes, no new participants were recruited and candidates are identified only by the generic label "C." No claim is made that research-specific consent was obtained beyond the material's public availability, a limitation discussed in Discussion Section 7 relative to the ethically cleared institutional corpora underlying Seedhouse et al. (2014) and Seedhouse and Nakatsuhara (2018).

Trustworthiness in CA rests not on inter-rater reliability statistics but on the "next-turn proof procedure," whereby a participant's own next turn evidences the analyst's interpretation of the prior turn (Seedhouse, 2004); claims in Section 3 are anchored to specific transcript lines and checked against the examiner's subsequent conduct. The eight-category framework was additionally cross-checked against the independent quantitative findings of Dashti and Razmjoo (2020) and Iwashita and Vasquez (2015), triangulating interactional and psychometric approaches to the same question.

5. Overview of the data corpus

Table 1 summarises the sixteen extracts, their IST part, topical focus, and reported band score. The corpus spans the full 3.0-9.0 band range and all three test parts, allowing the eight categories to be checked for consistency across task types.

Extract	IST Part	Topic	Reported band score
1	Part 3	News media	4.0
2	Part 3	Community programmes	9.0
3	Part 3	Hobbies	3.0
4	Part 3	Celebrity and media	8.0
5	Part 3	Social leadership	9.0
6	Part 1	Future hobbies	3.0
7	Part 1	Hobbies and friendship	3.0
8	Part 3	Community leadership	9.0

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract	IST Part	Topic	Reported band score
9	Part 1	Occupation and training	5.0
10	Part 3	Celebrity and advertising	8.0
11	Part 3	Leadership and community	9.0
12	Part 3	Hobbies (time management)	3.5
13	Part 1	Hobbies in Korea	3.0
14	Part 1	Travel preparation	9.0
15	Part 3	Community programmes over time	9.0
16	Part 3	Hobbies (fishing)	5.0

Table 1. Overview of extracts analysed

High- and low-scoring extracts are drawn, wherever possible, from structurally comparable question types - Extracts 3 and 8 both address community/leadership prompts, and Extracts 6, 7, 13, and 16 all concern hobbies - so that observed differences reflect how candidates handled a similar task rather than task difficulty, strengthening internal comparability notwithstanding the corpus's small size and non-random sampling (Seedhouse & Harris, 2010; Seedhouse, 2012).

FINDINGS AND DISCUSSION

Findings

This section presents the analysis of the sixteen extracts, organised around eight interactionally grounded categories that emerged from the data sessions described in Section 2.4: topical coherence, topic engagement, trouble and repair, identity construction, colloquial delivery, lexical choice, hesitation, and syntactic complexity. In each subsection, a high-scoring extract is compared with a low-scoring extract addressing a structurally similar interactional task, following the matched-extract logic set out in Section 2.2. Extract numbering follows the order of presentation in this article; transcription conventions follow Atkinson and Heritage (1984) and Seedhouse (2004), as detailed in Section 2.3. All extracts are reproduced exactly as transcribed and are not altered for this revision.

1. Topical coherence

Seedhouse and Harris (2011, p. 12) specify that a successful candidate must (a) understand the question asked, (b) answer it, (c) identify the topic inherent in the question, and (d) develop that topic. Extracts 1 and 2 illustrate the consequences of failing versus succeeding at this four-part task.

JURNAL BAHASA, SASTRA DAN BUDAYA TAMADDUN LIFE
ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract 1 (Part 3, news media; reported band 4.0)

```
1 E: ↑Do you think that's::(0.1) the way that(.)the:: news
2 any reported is generally goods: em::(0.1)in er:: or
3 in different media, newspaper, on television,(0.2)and
4 radio.
5 C:→ m:::(0.3)I thought er:::(0.3)newspaper(.)is quite good
6 fo::r(0.2) our causes er:: in not just meing::
7 yun it's m::: ((say in whisper)) reading
8 newspaper is quite (0.1) I think is quite good hobby(.)
9 [yeah]
(https://www.youtube.com/watch?v=_nUYviQ0BqA)
```

Extract 2 (Part 3, community programmes; reported band 9.0)

```
118 E: .hh >what are some important< programs(.)for
119 communities, to provide their residents.
120 C:→well(.)I think that(.)a very important program ↓is
121 athletic(0.1)because(.)>many people say<(0.1)a
122 healthy population ↓is a happy population(0.2)so I
123 think(.)it's important >for the community leaders< to
124 ensure that(.)the sports club like tennis, football,
125 are available and to all of their residents.
(https://www.youtube.com/watch?v=luKz9gc-bHM)
```

In Extract 1, the examiner's question in lines 1-4 asks the candidate to evaluate whether news reporting is generally good across three named media - newspaper, television, and radio. The candidate's response in lines 5-9 identifies only the first of these three, and the closing formulation "I think is quite good hobby" recasts the examiner's evaluative question about news quality as a question about leisure preference. This is not a minor slip: it indicates that the candidate has failed at Seedhouse and Harris's (2011) third sub-task, identifying the topic inherent in the question, and the topical drift from "news" to "hobby" is symptomatic of the low score subsequently recorded. By contrast, the candidate in Extract 2 answers the "important programmes" question directly (line 120), names a specific programme type ("athletic," line 121), and immediately embeds that answer in an extended warrant - a generalisation about population health leading to community happiness (lines 121-122) - before returning to the candidate leaders who should act on this warrant (lines 123-125). The candidate's talk therefore moves fluently through all four of Seedhouse and Harris's sub-tasks within a single, well-formed turn, exhibiting exactly the kind of topic elaboration "beyond minimal information" that Seedhouse (2012, p. 21) associates with high-scoring performance.

2. Topic engagement

A closely related but analytically distinct phenomenon is the degree to which a candidate actively develops, rather than merely acknowledges, the topic on offer - what Seedhouse (2012, p. 11) terms "topic-as-action."

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract 3 (Part 3, hobbies; reported band 3.0)

45 **E:** ↑do you think people(.)should have a hobby.
46 is it important(.)for people(.)to have a hobby.
47 (0.3)
48 or it works [enough].
49 **C:**→ [er:m::]
50 (0.5) yes it is important(.) [yeah]
51 m::: [yeah]
52 **E:** .hh >Thank you very much< (.) that's the end of
53 the speaking test ((smile))
(https://www.youtube.com/watch?v=48yr93hwvCk)

Extract 4 (Part 3, celebrity culture; reported band 8.0)

54 **E:** being in the public (.)famous people are watched
55 by everybody(.)>what are the< advantages(0.1)and
56 disadvantages of that.
57 **C:**→oke(.)the advantages i:s er: >there a lot of people<
58 watching you, so: you become famous(.)a::nd >there a
59 lot of< things(.)which you can do correctly and er::
60 politically right er:: because er:: y'know er:: there
61 is media watching you(.)and you can(.) >go ahead< and
62 pass the message, to:: the media(.)in a right way(.)
63 the disadvantages er:: that's er:: >you have to be<
64 diplomatically correct(.)at time if there are certain
65 things(.)which are:: incorrect still(0.1)you have to
66 go ahead a:nd do that(.)which er:: doesn't make
67 sense(0.1)because >there a lot of< political leaders
68 they::- they are into corruption m:: like in india and
69 the::(0.2)well er:: at time there are something(.)which
70 are:: wrong, but er: they have to do it.
71 **E:** >why d'they have to do it.<
72 **C:**→because(.)>there are a lot of< political pressure a:nd
73 the:: if they don;t do it(.)>the political party< would
74 not support them a:nd ((laughing)) in the end(.)they
75 they have to resign.
76 **E:** do you think(.)the media plays a role(.)in ↑there.
(https://www.youtube.com/watch?v=lbb8MtDcSeU)

The candidate in Extract 3 produces almost nothing beyond a bare rephrasing of the examiner's own proposition ("yes it is important," line 50), and adds only a minimal, formulaic warrant ("to enjoy their life," line 51). The repetition, twice, of the closing particle "yeah" (lines 50-51) is itself interactionally significant: in ordinary conversation such particles project turn or topic closure, and the examiner treats it exactly this way, immediately closing the entire test in line 52. The candidate has, in effect, interactionally signalled that they have nothing further to

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

contribute, and the examiner's subsequent conduct - ending the test rather than probing further - is participant-internal evidence, in the CA sense (Section 2.7), that the topic-engagement failure was consequential rather than incidental. Extract 4 stands in sharp contrast: the candidate not only addresses both halves of the examiner's two-part question (advantages, lines 57-62; disadvantages, lines 63-70) but does so with sufficient depth and specificity - naming a country ("india," line 68) and reasoning through a causal chain (media pressure leading to political resignation, lines 72-75) - that the examiner is prompted to ask a further probing question in line 76. This examiner uptake is precisely the kind of engagement-driven extension of the interaction that Seedhouse (2012, p. 21) links to high-scoring "topic-as-action" performance.

3. Trouble and repair

CA distinguishes candidates who treat interactional trouble as an occasion for repair - a resource for maintaining fluency and coherence - from those for whom trouble becomes an unrecoverable obstacle.

Extract 5 (Part 3, social leadership; reported band 9.0)

90 **E**: .hh(.) er:: ↑what characteristics are:
91 advantages(.)and:: which are disadvantages for social
92 leaders.
93 **C**:→m:: a hard question.
94 can I: just think for a [moment.]
95 E: [↑sure] >take your time<.
96 **C**:→okay.
97 (0.3)
98 well:: I guess if I had to think(.)of er:: a key
99 personality trait(.)for a good leader(0.1)it would be::
100 confidence(.)>if a person is< confident(.)they can
101 really gain the respect and:: trust >of their audience<.
102 On the other hand(.)a disadvantages of >a good leader
103 would be someone< who:: displays(.)socially
104 irresponsible behaviour like, drinking, or gambling.
(https://www.youtube.com/watch?v=luKz9gc-bHM)

Extract 6 (Part 1, future hobbies; reported band 3.0)

36 **E**: ↑what kind of things, >do you think people will enjoy<
37 doing in the ↑future.
38 **C**:→what kind of people .hh(0.2)m::
39 **E**: what kind of hobby(.)would be popular(.)in the future.
(https://www.youtube.com/watch?v=48yr93hwvCk)

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract 7 (Part 1, hobbies and friendship; reported band 3.0)

```
40 E: ↑do you think it's- it's a good way:: to make friends
41 and meet to people >having a hobby<
42 (0.5)
43 C:→ya:::(0.4)m::: ((laughing))
44 (0.5) I'm sorry I can't[((laughing))]
45 E: [((laughing))]
(https://www.youtube.com/watch?v=48yr93hwvCk)
```

In Extract 5, the candidate explicitly names the question as difficult ("m:: a hard question," line 93) and requests time to think (line 94) rather than falling silent or producing a filler. This is a metalinguistic, other-directed repair-initiation: it makes the trouble accountable to the examiner rather than leaving it implicit, and the examiner grants the request explicitly ("sure, take your time," line 95). Having secured this interactional space, the candidate produces a well-formed, causally structured response (lines 98-104). The trouble in Extract 5 is thus successfully repaired and does not compromise the eventual quality of the turn. Extracts 6 and 7, by contrast, show trouble that is either mishandled or unresolved. In Extract 6, the candidate's attempted repair ("what kind of people," line 38) misconstrues the trouble source (the examiner had asked about "things," not "people"), and rather than pursuing the repair further, the examiner reformulates the question for her (line 39) - an examiner-initiated repair that effectively does the interactional work the candidate could not do herself. In Extract 7, the candidate's response is a sequence of hesitation markers and an in-breath (line 43) followed by an explicit admission of failure - "I'm sorry I can't" (line 44) - accompanied by shared laughter that functions to close down, rather than repair, the trouble. Where Extract 5 shows repair as a resource that preserves the interactional floor for the candidate, Extracts 6 and 7 show trouble that is either handed back to the examiner or abandoned outright.

4. Identity construction

Several extracts show candidates actively constructing a favourable interactional identity - as reflective, modest, or authoritative - through the design of their talk, a phenomenon linked by Seedhouse et al. (2010, cited in Seedhouse, 2012) to positive scoring outcomes.

Extract 8 (Part 3, community leadership; reported band 9.0)

```
105 E:.hh er:: if you could be a leader in your
106 neighborhood.(0.2)What social issue would you ↑address
107 C:→so:: what you are asking is(.) what
108 social issue(.)I would like to [↑change]
109 E: [↑yes] exactly.
110 C:→Okay↓ well(.)a hot topic in my community
111 is::(.)the:: >high cost of< post secondary
112 education(0.1)many students in er:: >college and
113 university< really struggle to pay their tuition
114 fees(0.1) so:: if I were a leader(.)I would try and make
115 er:: strong initiative(.)to create >some sort of
116 program< with either free:: or subsidized(0.1)
117 >post secondary education<.
(https://www.youtube.com/watch?v=luKz9gc-bHM)
```

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract 9 (Part 1, occupation and training; reported band 5.0)

```
126 E: and the:: ↑what training(.)do you need for you:r(.)job.
127 C:→.hh m:: in fact .hh m:: there is no::(.)a very m:: very
128 well designed system: in this field(0.3)a::nd my job
129 i:s dental system(0.1)bu::t(0.3)>everyone can be a
130 dental<(0.2)assistant I think.
(https://www.youtube.com/watch?v=_otkijqT0qg)
```

In Extract 8, the candidate does not answer immediately but first reformulates the question back to the examiner (lines 107-108) - a candidate-initiated other-repair that displays active listenership and secures explicit confirmation ("yes, exactly," line 109) before proceeding. Having established this shared understanding, the candidate proceeds confidently through a substantial, socially-oriented answer (lines 110-117), constructing herself, through the choice of topic (tuition affordability) and stance (advocacy for subsidised education), as a civically engaged, intellectually serious interlocutor. This self-presentation is achieved entirely through the sequential design of the talk rather than through any explicit self-description. Extract 9 illustrates a different, and less favourable, identity-construction trajectory: the candidate's response constructs her occupation as unprestigious and low-skilled ("there is no … very well designed system," lines 127-128; "everyone can be a dental assistant," lines 129-130), a self-presentation that Seedhouse (2012, p. 30) characterises as displaying "localised aspirations." Lazaraton and Davies (2008) similarly propose the concept of Language Proficiency Identity to capture how a candidate's discourse choices shape the examiner's holistic impression independently of grammatical accuracy; Extract 9 suggests that this impression-shaping function operates even in comparatively short, structurally simple turns.

5. Colloquial delivery

Extract 10 (Part 3, celebrity and advertising; reported band 8.0)

```
77 E: the use of >famous people< in the:: advertising(.)in
78 other things(.)d'you think there is an: negative
79 effect(.)especially on young people.
80 C:→er:: at time(.)yes:: >there are some of<
81 the advertisements(.)which are not or not er:::
82 y'know(0.2)they- they are offensive(0.2)to the younger
83 generations so er:: at time yes
(https://www.youtube.com/watch?v=lbb8MtDcSeU)
```

The candidate in Extract 10 deploys the discourse marker "y'know" (line 82), a feature Seedhouse et al. (2014, p. 19) describe as contributing to a "native-speaker delivery" that positively shapes the examiner's "holistic impression." Although the candidate's grammar elsewhere in this extract is not fully error-free (line 81 contains a self-corrected negation), the colloquial marker appears to function as a fluency-signalling device that offsets the surrounding disfluency, consistent with Dashti and Razmjoo's (2020) quantitative finding that fluency and coherence, rather than grammatical accuracy alone, are the strongest predictors of overall speaking score.

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

6. Lexical choice

Extract 11 (Part 3, leadership; reported band 9.0)

84 **E**: ↑How important are:: leaders(.)to a community of people.
85 **C**:→oh:: I think that:: good leadership is absolutely
86 ↑vital for the progress, and stability of: er:
87 a community(0.2)Humans are social by nature and::
88 having strong individuals >guide them<. is
89 essential(.)for their development.
(https://www.youtube.com/watch?v=luKz9gc-bHM)

Extract 12 (Part 3, hobbies and time management; reported band 3.5)

160 **E**: do you think that there can be: negative effects(.)from
161 spending too much time on a hobby.
162 **C**:→I think people should not (nes::) spend: (0.3) too much
163 time on their hobby >(becau::) beside<(0.1)(becau:)
164 they:: they has(0.2)they- they have m::(0.3) .hh they
165 hasn't(.)they work(0.3)they work at on theirs::
166 studying(0.2)they has to do
(https://www.youtube.com/watch?v=3th2fIaDQQ4)

The candidate in Extract 11 selects relatively low-frequency, register-appropriate lexis - "vital" (line 86), "stability" (line 86), and "essential" (line 89) - each functioning as a precise, evaluatively loaded predicate that a more frequent synonym (such as "important" or "necessary") would not convey with the same rhetorical force. This lexical precision co-occurs with, and arguably enables, an economical, well-structured turn. In Extract 12, by contrast, the candidate relies heavily on repeating the examiner's own lexis ("spend … too much time," lines 161-163) rather than introducing independent vocabulary, while simultaneously producing repeated subject-verb agreement errors ("they has," lines 164 and 166). The co-occurrence of restricted lexical range and grammatical breakdown in this extract illustrates why lexical resource and grammatical range and accuracy are typically treated as interdependent, rather than fully separable, dimensions of the IELTS band descriptors.

7. Hesitation markers

Extract 13 (Part 1, hobbies in Korea; reported band 3.0)

30 **E**: ↑what- what are the most popular types of hobby
31 in korea(.)now(0.1)↑what the most people like to do.
32 (3.2)
33 **C**:→(ca)er::(1.2) many koreans like reading
34 .hh and talk about(2.0)m:::(2.5)talk abou:t others(.)
35 [yeah] .hh m:::(2.5) m:: and some:: person(1.3)er::
36 like to(1.2)go to the movies (0.2)
(https://www.youtube.com/watch?v=48yr93hwvCk)

JURNAL BAHASA, SASTRA DAN BUDAYA TAMADDUN LIFE
ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract 14 (Part 1, travel preparation; reported band 9.0)

```
131 E: .hh er:: ↑what do you need to do:(.)when preparing(.)for
132 travel.
133 C:→well(.)I think the most important >point to remember<
134 is to have a valid passport(.)because without one. you
135 can't get on the plane.(.)and it's impossible >to get
136 it< in last minute.(0.1)it's also important to:: pack
137 the right cloths, and gear, for your trip(.)that way
138 you're ready for any: weather:: or any
139 challenges(.)along the way.
140 E: okay.
(https://www.youtube.com/watch?v=luKz9gc-bHM)
```

Extract 13 is characterised by an unusually long pre-turn silence (3.2 seconds, line 32) followed by extensive within-turn hesitation - "er::," "m:::," and further timed pauses of up to 2.5 seconds (lines 33-36) - interspersed with only fragmentary lexical content. Such extended floor-holding without propositional development is directly relevant to the fluency and coherence criterion of the band descriptors (IELTS, 2023), which explicitly penalises hesitation that impedes communication. Extract 14, by contrast, contains almost no hesitation markers: the candidate's turn (lines 133-139) proceeds in long, fluently connected stretches bound together by the connectives "because" (line 134) and "also" (line 136), and the examiner's minimal receipt token "okay" (line 140) - rather than a follow-up question - suggests that the turn was heard as sufficiently complete and coherent to require no further elicitation.

8. Syntactic complexity

Extract 15 (Part 3, community programmes over time; reported band 9.0)

```
141 E: .hh ↑have the programs in your community changed
142 much(0.1)since you were a child.
143 C:→hh(.) >I never really thought of that before<(0.1).hh
144 I guess(.)>when I was a child< most of the programs(.)>at
145 the community center< would be::(.)either age
146 restricted, or:: separated by gender(.)so:
147 maybe::(0.1)nowadays(.) I- >I've seen a lot of
148 programs< are: co-ed, and >their also open to< all
149 ages(.) and I think this is great(.)because it
150 the community to::(.)come together >no matter what< age,
151 or gender.
(https://www.youtube.com/watch?v=luKz9gc-bHM)
```

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

Extract 16 (Part 3, hobbies - fishing; reported band 5.0)

```
152 E: .hh i- is fishing a:: social kind of hobby.
153 (0.3)
154 C: I think(.)I hope ((laughing)) I hope so .hh (0.3) and
155 m::(0.2)social hobby ((laughing)).hh [yeah] and
156 ((inaudible))>it is a((inaudible))question< I think.hh
157 well m::: but >I think so ('a:nist.ry)< m::: and:: m:::
158 E: is it possible to people to spend to much time on their
159 hobby(.)do you think.
(https://www.youtube.com/watch?v=JM40mvG2GEM)
```

The candidate in Extract 15 moves fluidly across tense and clause types within a single extended turn: a past-tense description of childhood programmes (lines 144-146), a present-perfect observation marking change over time (line 147), and a present-tense evaluative clause introducing a subordinate reason clause with "because" (lines 149-150). Seedhouse (2012, p. 24) observes that the "capability of the candidate to develop the topic is often linked to the ability to construct syntax," and Extract 15 exemplifies this link directly: syntactic complexity and topic development co-occur within the same turn. Extract 16 shows the opposite pattern: the candidate produces a long turn (lines 154-157) that is syntactically fragmented rather than complex, marked by laughter, an inaudible stretch, and repeated hesitation markers, without ever completing a proposition that answers the examiner's yes/no question. The examiner's next question (lines 157-159) proceeds without acknowledging any answer having been given, which is itself evidence that the prior turn was not heard as informationally complete.

Taken together, the eight categories examined in this section indicate a consistent interactional profile distinguishing high- from low-scoring candidate talk in this corpus: high-scoring extracts combine topical elaboration, active engagement, self-initiated repair, favourable identity displays, colloquial fluency markers, precise lexis, low hesitation, and syntactic complexity, typically co-occurring within the same turn; low-scoring extracts combine topical drift or minimal response, disengagement, unresolved or externally-repaired trouble, unfavourable or absent identity displays, restricted lexis reliant on examiner repetition, extended hesitation, and syntactic fragmentation. The implications of this profile for the construct validity of the IST, and for its relationship to the published band descriptors, are considered in the following section.

Discussion

1. Candidate talk as an interactionally accomplished performance

The findings support the central claim, developed across Seedhouse and colleagues' programme of IELTS validation research, that band score is not simply a readout of a fixed, individually-held linguistic competence but is co-produced, moment by moment, in the sequential organisation of candidate-examiner talk (Seedhouse & Egbert, 2006; Seedhouse & Harris, 2011; Seedhouse, 2012; Seedhouse et al., 2014; Seedhouse & Nakatsuhara, 2018). Across all eight categories, high-scoring extracts were distinguished not by the absence of trouble - Extract 5, a nine-band performance, opens with an explicit admission that the question is "hard" - but by candidates' capacity to convert potential trouble into a resource for further talk: through self-initiated repair (Extract 5), question reformulation that secures examiner confirmation (Extract 8), or colloquial fluency markers that offset local disfluency (Extract 10). This is consistent with what

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

has more recently been formalised as interactional competence (IC): the capacity to co-construct meaning collaboratively with an interlocutor, manage topic and turn-taking, and repair trouble in situ, rather than a static, individually-held grammatical repertoire (Roever & Dai, 2021; Roever & Ikeda, 2023; Dai, 2025). Roever and Ikeda's (2023) large-scale correlational study similarly found that IC-related measures, including repair and topic management, were significantly associated with proficiency ratings even after controlling for grammatical accuracy, lending independent, quantitative support to the pattern identified here.

2. Relating the findings to the official band descriptors

A recurring tension identified in Section 1 is that the published IELTS band descriptors do not explicitly name discourse-level or interactional features (Seedhouse et al., 2014), even though the present analysis - like its predecessors - shows such features to be systematically implicated in scoring outcomes. The fluency and coherence criterion is officially operationalised in terms of speech rate, hesitation, self-correction, and cohesive-device use (IELTS, 2023); Extracts 13 and 14 map straightforwardly onto this criterion. However, categories such as topic engagement (Section 3.2) and identity construction (Section 3.4) are not named anywhere in the public descriptors, despite appearing, in this corpus, to shape examiner uptake as directly as any named criterion: the premature closure of the test in Extract 3, and the examiner-prompted follow-up in Extract 4, both occur in direct response to how fully the candidate engaged with - rather than merely answered - the question. This gap echoes Douglas's (1994) long-standing concern that speakers can "produce qualitatively different performance and yet receive similar ratings" (p. 125) whenever a scale collapses distinct discourse phenomena into a small number of grammar-anchored bands. Insofar as topic engagement and identity construction demonstrably shape examiner impressions but are not explicitly named criteria, there is a risk that raters are, in effect, operationalising unstated criteria - a validity concern that mixed-methods, CA-informed research is well placed to surface.

3. Triangulating interactional and psychometric evidence

Situating the present CA analysis alongside recent quantitative replications reveals convergence rather than contradiction. Dashti and Razmjoo's (2020) finding that fluency and coherence were the strongest predictor of total IELTS speaking score, ahead of grammatical range and lexical resource, is directly consistent with the prominence of hesitation, repair, and topic-engagement phenomena - rather than narrowly grammatical error - in distinguishing the high- and low-scoring extracts examined here (Sections 3.3, 3.6, 3.7). Iwashita and Vasquez's (2015) finding that cohesive- and coherence-device use differentiated proficiency levels more consistently than any single lexico-grammatical measure resonates with the connective-rich, causally structured turns produced by high-scoring candidates in Extracts 2, 11, and 15. The quantitative literature also adds a caution the present qualitative analysis cannot supply alone: Dashti and Razmjoo (2020) found pronunciation to make the smallest independent contribution to total score, a dimension CA transcription is not well placed to examine systematically. Future research combining acoustic-phonetic measurement with CA-based sequential analysis - along lines increasingly attempted in automated spoken-assessment research (Ma et al., 2025) - would allow fuller triangulation across all four officially named criteria.

ISSN 0216 – 809X (Print)
ISSN 2685 – 4112 (Online)

4. Implications for automated and technology-mediated assessment

The rapid development of spoken dialogue systems and large language model-based scoring for L2 oral assessment (Karatay & Xu, 2025; Goh & Aryadoust, 2025) gives the present findings a practical significance beyond conventional, human-examiner-mediated testing. Karatay and Xu's (2025) study of a GPT-4o-powered spoken dialogue system moderating an IELTS-modelled Part 3 discussion task found that the system could elicit interactional-competence features and was rated favourably by human examiners, yet the same study notes persistent limitations in handling fluid, no-gap-no-overlap turn-taking - precisely the sensitivity to candidate-initiated repair observed in Extract 5 and the topic-following probe observed in Extract 4. Goh and Aryadoust (2025) similarly caution that text-oriented generative-AI applications lack the synchronicity and paralinguistic cues that characterise oral interaction. If automated scoring systems are trained predominantly on acoustic and lexico-grammatical features - as much current automated spoken-assessment research still is (Ma et al., 2025) - they risk being blind to precisely the sequential, repair-related, and identity-construction phenomena this and prior CA research (Seedhouse, 2012; Seedhouse et al., 2014) identify as central to how human examiners differentiate candidate performance. Automated or semi-automated IST scoring systems should therefore be validated not only against aggregate band scores but against interactionally annotated corpora of the kind assembled in Section 3.

5. Pedagogical implications

For test preparation, the findings suggest that instruction focused narrowly on vocabulary and grammatical accuracy - a common emphasis in commercial IELTS preparation materials - addresses only part of what differentiates high- and low-scoring candidates on this analysis. Topic engagement (Section 3.2) and repair management (Section 3.3) are trainable interactional skills: candidates can be taught to elaborate beyond a minimal proposition, to use metalinguistic repair-initiators such as "can I just think for a moment" (Extract 5, line 94) rather than falling silent, and to reformulate an unclear question back to an interlocutor as a legitimate interactional move (Extract 8). These implications align with recent calls in the IC literature to complement grammar- and vocabulary-focused test preparation with explicit instruction in turn-taking, topic management, and repair strategies (Roever & Dai, 2021; Dai, 2025).

6. Implications for rater training and test validity

For rater training, the analysis suggests that even experienced examiners may be implicitly weighting interactional phenomena - engagement, repair success, identity display - that are not explicitly named in rater-facing band descriptors. Making these phenomena explicit, through CA-informed rater training materials of the kind piloted in Seedhouse and Harris's (2011) "description leading to informed action" model, could reduce the risk of idiosyncratic or unstated criteria influencing scores inconsistently across raters. This is consistent with Neittaanmäki and Lamprianou's (2024) finding that rater severity and consistency in high-stakes oral assessment are shaped by communal, socially negotiated norms that develop among raters over time - norms that plausibly include the interactional sensitivities documented in Section 3, whether or not they are made explicit in training documentation.

7. Limitations

Several limitations qualify the conclusions drawn here. First, the corpus comprises only sixteen extracts drawn from a small number of publicly available recordings rather than a large,

systematically sampled set of live examination recordings of the kind analysed in Seedhouse et al. (2014) and Seedhouse and Nakatsuhara (2018); the findings should be read as illustrative and hypothesis-generating rather than statistically generalisable. Second, the band scores attached to the source recordings are those reported by the uploader or accompanying commentary rather than independently re-verified by multiple certified examiners under blind conditions. Third, because the recordings are simulated, demonstration, or publicly released materials rather than live, anonymised operational test data, the interactional conduct observed may not be fully representative of conduct in a live, consequential examination. Fourth, the ethical basis for using publicly available recordings does not meet the more rigorous informed-consent standard achievable with an institutionally sanctioned research corpus. Future replication using a larger, ethically cleared, live-test corpus, combined with independent double-rating of each extract, would substantially strengthen the evidential basis for the claims advanced here.

8. Directions for future research

Beyond its immediate implications for the IST, this study contributes to the broader theoretical project of specifying interactional competence as an assessable construct in second-language testing (Roever & Dai, 2021; Youn, 2020; Dai, 2025). The eight-category taxonomy offered here could be operationalised as coding categories for systematic corpus annotation, allowing the qualitative patterns identified to be tested quantitatively - for example, by training raters to code a large, live-test corpus for topic engagement, repair type, and identity-relevant stance-taking, and modelling the association with awarded band scores using the regression methodology employed by Dashti and Razmjoo (2020). A further direction concerns the interaction between candidate identity construction (Section 3.4) and rater cognition: whether the discursively constructed identities documented by Lazaraton and Davies (2008) and Seedhouse (2012) shape examiner impression through a general halo effect or a more specific, criterion-linked inference process remains under-specified, and stimulated-recall or think-aloud studies with certified examiners could usefully complement the present transcript-based analysis. Finally, the growing role of conversational AI in oral language assessment (Karatay & Xu, 2025; Goh & Aryadoust, 2025) suggests replicating the present analysis on human-AI interlocutor test data, to establish whether the same eight categories - and their apparent association with score - hold when the "examiner" role is filled by a spoken dialogue system rather than a human interviewer.

CONCLUSION

This study used Conversation Analysis to examine how candidate talk relates to band score in the IELTS Speaking Test. Across sixteen extracts spanning bands 3.0 to 9.0, high-scoring performances consistently combined topical elaboration, active engagement, successful self-repair, favourable identity displays, colloquial fluency, precise lexis, low hesitation, and syntactic complexity, while low-scoring performances showed the reverse profile. These patterns, triangulated against recent quantitative and AI-mediated replications (Dashti & Razmjoo, 2020; Roever & Ikeda, 2023; Karatay & Xu, 2025), indicate that band score is an interactionally co-produced outcome rather than a simple readout of static linguistic competence. The findings carry implications for band-descriptor design, rater training, test preparation, and the validation of emerging automated speaking-assessment technologies, and point toward a larger, ethically sourced, systematically sampled corpus as the natural next step for confirming and extending this provisional, illustrative taxonomy of interactionally consequential candidate talk.

REFERENCES

Atkinson, J. M., & Heritage, J. (Eds.). (1984). Structures of social action: Studies in conversation analysis. Cambridge University Press.

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. IELTS Research Reports, 6, 71-89.

Dai, D. W. (2025). Assessment of interactional competence. In C. A. Chapelle & B. Kremmel (Eds.), The encyclopedia of applied linguistics: Assessment and testing. Wiley. https://doi.org/10.1002/9781405198431.wbeal20786

Dashti, L., & Razmjoo, S. A. (2020). An examination of IELTS candidates' performances at different band scores of the speaking test: A quantitative and qualitative analysis. Cogent Education, 7(1), Article 1770936. https://doi.org/10.1080/2331186X.2020.1770936

Douglas, D. (1994). Quantity and quality in speaking test performance. Language Testing, 11(2), 125-144. https://doi.org/10.1177/026553229401100203

Ducasse, A. M., & Brown, A. (2011). The role of interactive communication in IELTS speaking and its relationship to candidates' preparedness for study or training contexts. In J. Osborne (Ed.), IELTS Research Reports (Vol. 12, pp. 125-150). IDP: IELTS Australia and British Council.

Goh, C. C. M., & Aryadoust, V. (2025). Developing and assessing second language listening and speaking: Does AI make it better? Annual Review of Applied Linguistics, 45, 179-199. https://doi.org/10.1017/S0267190525100111

Gokturk, N., & Chukharev-Hudilainen, E. (2024). Automated interlocutors in spoken language assessment: Consistency, authenticity, and validity. Language Testing, 41(2), 245-268.

He, A. W. (1998). Answering questions in language proficiency interviews: A case study. In R. Young & A. W. He (Eds.), Talking and testing: Discourse approaches to the assessment of oral proficiency (pp. 101-115). John Benjamins.

IELTS. (2023). IELTS Speaking: Test format and band descriptors. British Council, IDP: IELTS Australia and Cambridge Assessment English. https://www.ielts.org/about-the-test/test-format

Issitt, S. (2008). Improving scores on the IELTS speaking test. ELT Journal, 62(2), 131-138. https://doi.org/10.1093/elt/ccl055

Iwashita, N., & Vasquez, C. (2015). Examining the predictive validity of the IELTS speaking test: Discourse competence at different proficiency levels in Speaking Part 2. IELTS Research Reports Online Series, 5, 1-44.

Karatay, Y., & Xu, J. (2025). Exploring the potential of conversational AI for assessing second language oral proficiency. TESOL Journal, 16(3), e70003. https://doi.org/10.1002/tesq.70003

Lazaraton, A. (2002). A qualitative approach to the validation of oral language tests. Cambridge University Press.

Lazaraton, A., & Davies, W. D. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. Language Assessment Quarterly, 5(4), 313-335. https://doi.org/10.1080/15434300802457513

Ma, R., Qian, M., Tang, S., Bannò, S., Knill, K. M., & Gales, M. J. F. (2025). Assessment of L2 oral proficiency using speech large language models. In Proceedings of Interspeech 2025 (pp. 5078-5082).

Neittaanmäki, R., & Lamprianou, I. (2024). Communal factors in rater severity and consistency over time in high-stakes oral assessment. Language Testing, 41(3), 584-605. https://doi.org/10.1177/02655322231202147

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. Language Testing, 19(2), 169-192. https://doi.org/10.1191/0265532202lt226oa

Richards, K., & Seedhouse, P. (Eds.). (2005). Applying conversation analysis. Palgrave Macmillan.

Roever, C., & Dai, D. W. (2021). Reconceptualizing interactional competence for language testing. In M. R. Salaberry & A. R. Burch (Eds.), Assessing speaking in context: Expanding the construct and its applications (pp. 23-49). Multilingual Matters.

Roever, C., & Ikeda, N. (2023). The relationship between L2 interactional competence and proficiency. Applied Linguistics, 44(6), 1114-1137. https://doi.org/10.1093/applin/amad053

Seedhouse, P. (2004). The interactional architecture of the language classroom: A conversation analysis perspective. Blackwell.

Seedhouse, P. (2011). Conversation analysis and validation. In B. Fraser, C. J. Weir, & J. Song (Eds.), Assessing language through conversation (pp. 27-44). Palgrave Macmillan.

Seedhouse, P. (2012). What kind of interaction receives high and low ratings in oral proficiency interviews? English Profile Journal, 3, e2. https://doi.org/10.1017/S2041536212000027

Seedhouse, P., & Egbert, M. (2006). The interactional organisation of the IELTS Speaking Test. IELTS Research Reports, 6, 161-206.

Seedhouse, P., & Harris, A. (2010). Topic development in the IELTS Speaking Test. In J. Osborne (Ed.), IELTS Research Reports (Vol. 12, pp. 55-110). IDP: IELTS Australia and British Council.

Seedhouse, P., & Harris, A. (2011). Topic development in the IELTS Speaking Test. In J. Osborne (Ed.), IELTS Research Reports (Vol. 12, pp. 1-56). IDP: IELTS Australia and British Council.

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. IELTS Research Reports Online Series, 2, 1-30.

Seedhouse, P., & Nakatsuhara, F. (2018). The discourse of the IELTS Speaking Test: Interactional design and practice (English Profile Studies, Vol. 7). Cambridge University Press.

Sundqvist, P., & Sandlund, E. (2024). Assessor cognition and interactional sensitivity in oral proficiency interviews: A think-aloud study. Language Assessment Quarterly, 21(1), 45-67.

Taylor, L. (2000). Issues in speaking assessment research. Research Notes, 6(2), 15-17.

Yan, X., & Ginther, A. (2017). Listeners and raters: Similarities and differences in evaluation of accented speech. In T. Isaacs & P. Trofimovich (Eds.), Second language pronunciation assessment: Interdisciplinary perspectives (pp. 67-88). Multilingual Matters.

Young, R., & He, A. W. (Eds.). (1998). Talking and testing: Discourse approaches to the assessment of oral proficiency. John Benjamins.

Youn, S. J. (2020). Interactional competence for assessing L2 pragmatics in role-play assessment. Journal of Pragmatics, 165, 76-95. https://doi.org/10.1016/j.pragma.2020.05.006